

Deep Neural Profiling Identifies RAP1GAP2 as a Latent Regulator in Head and Neck Squamous Cell Carcinoma: Integrative Transcriptomic and Multi-omic Analysis

Joy Prokash Debnath¹, Tanvir Hossain^{1*}, Papia Rahman^{1*}

¹Department of Biochemistry and Molecular Biology, School of Life Science, Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

*Corresponding authors: Tanvir Hossain (tanvir-bmb@sust.edu); Papia Rahman

Preprint, version 2 (October 2026)

¹ Abstract

² **Background:** Conventional differential gene expression (DGE) analysis inadequately captures the
³ complex molecular changes that drive the progression of head and neck squamous cell carcinoma
⁴ (HNSCC). Variational autoencoders (VAEs) offer a deep learning approach to uncover hidden patterns in
⁵ high-dimensional transcriptomic data, and the DeepProfile framework showed that combining VAEs with
⁶ Integrated Gradients yields interpretable latent spaces across human cancers.

⁷ **Methods:** Following DeepProfile, the head and neck cancer expression compendium (643 samples from
⁸ 26 GEO datasets, 11,020 genes) was compressed to 500 principal components and used to train a VAE
⁹ with 50 latent variables. Integrated Gradients was used to determine the contribution of each gene to each
¹⁰ latent variable. Genes with consistently high attribution across latent variables were selected as candidate
¹¹ regulators and characterised by pathway enrichment, GSEA, supervised deep learning classifiers and
¹² DGE analysis in independent RNA-seq data. The top candidate, RAP1GAP2, was further examined
¹³ across DNA methylation, copy number, mutation, paired tumour–normal expression and survival in
¹⁴ TCGA-HNSC and GSE178537, together with a genome-wide integration of promoter methylation and
¹⁵ gene expression.

¹⁶ **Results:** The VAE latent space captured distinct gene programs and pathways. RAP1GAP2 was among
¹⁷ the 20 genes with the highest mean attribution and was the most important feature in the supervised
¹⁸ classifiers: an MLP trained on the 18 measured candidate genes reached a mean AUPRC of 0.86 and
¹⁹ AUROC of 0.80, and a RAP1GAP2-only classifier reached a mean AUPRC of 0.769. This occurred
²⁰ despite the lack of substantial differential expression in tumours relative to normal samples, which was
²¹ confirmed in 43 paired TCGA-HNSC patients ($\log_2\text{FC} -0.13$, adjusted $p = 0.57$) and 20 paired
²² GSE178537 patients ($\log_2\text{FC} +0.31$, adjusted $p = 0.17$). The RAP1GAP2 promoter was hypomethylated
²³ in tumours (7 of 10 promoter CpGs; mean $\Delta\beta -0.20$), but methylation was only weakly associated with
²⁴ expression within tumours (Spearman $\rho = -0.26$). RAP1GAP2 was rarely mutated (1/515), showed only
²⁵ shallow copy-number changes and was not associated with overall survival (HR 1.02, $p = 0.91$).
²⁶ Genome-wide, 163 hypermethylated-downregulated and 42 hypomethylated-upregulated genes showed a
²⁷ significant negative methylation–expression correlation.

²⁸ **Conclusion:** Our deep learning framework identified RAP1GAP2 as a latent, classification-informative
²⁹ gene in HNSCC that is overlooked by DGE analysis. Its contribution is not explained by mutation, copy
³⁰ number or promoter methylation, and biological interpretation suggests a possible role through Rap1,
³¹ MAPK signalling and Golgi-mediated secretion that now requires experimental validation.

Keywords: Head and neck squamous cell carcinoma (HNSCC); transcriptomics; deep learning; variational autoencoder; DeepProfile; Integrated Gradients; RAP1GAP2; DNA methylation; Rap1 signalling

1. Introduction

1.1 Overview of head and neck squamous cell carcinoma

Head and neck cancers arise in the oral cavity, pharynx and larynx (Figure 1A), and most are head and neck squamous cell carcinomas (HNSCC) (Johnson et al., 2020). HNSCC is of particular clinical significance because human papillomavirus (HPV) infection has driven a rise in its incidence in recent decades, mainly in tumours of the oropharynx, making HPV-positive HNSCC one of the fastest-growing cancers in many high-income nations (Lechner et al., 2022). Because of subtle early symptoms, these cancers frequently present at advanced stages, leading to substantial morbidity and mortality. A deeper understanding of their molecular basis is needed for earlier detection, better patient stratification and more effective precision therapies, and a better understanding of the tumour transcriptome may help identify molecular drivers that could enhance patient care. Large-scale genomic profiling by The Cancer Genome Atlas has defined the mutational and transcriptional landscape of HNSCC (Cancer Genome Atlas Network, 2015).

1.2 Complexity of cancer biology and analytical gaps

The biology of cancer involves non-linear interactions among genes and pathways. Traditional differential gene expression (DGE) analysis, which relies on linear models or statistical tests to find genes individually up- or down-regulated in tumours, excels at identifying genes with large average expression changes but may overlook genes that act through subtle or combinatorial patterns. Patient subgroups or tumour phenotypes could therefore be determined by gene sets that show no obvious one-at-a-time differences and remain invisible to linear DGE approaches (Rampášek et al., 2019; Sabbatini and Manganaro, 2023). Explainable machine learning, which applies interpretability techniques to complex models, can reveal such multivariate gene patterns (Abbas and El-Manzalawy, 2020; Way et al., 2020).

1.3 Deep learning for latent feature discovery

Unsupervised deep neural networks can learn biologically meaningful latent variables from transcriptomic data. Variational autoencoders (VAEs) (Kingma and Welling, 2014) are deep generative models that compress high-dimensional gene expression profiles into a lower-dimensional latent space while preserving as much information as possible, reducing complex expression patterns to latent features that capture patient variability and underlying biological signals (Figure 1B). Rampášek et al. showed that a VAE-based model (Dr.VAE) learns latent representations of cancer cell line expression that improve drug response prediction (Rampášek et al., 2019), and Way et al. found that compressing pan-cancer expression with different latent dimensionalities reveals complementary biological signals (Way et al., 2020). DeepProfile extended this approach to 50,211 transcriptomes across 18 human cancers and showed that its latent spaces are more biologically interpretable than other dimensionality-reduction methods (Qiu et al., 2025). A known drawback of deep learning models is limited interpretability: compressed features that correlate with disease cannot be translated into testable biological insights without knowing which genes they represent.

72 1.4 Gene attribution with Integrated Gradients

73 Integrated Gradients (IG) quantifies the contribution of each input feature, here each gene's expression,
74 to a given output or latent variable by integrating the model's gradients along a path from a baseline to
75 the actual input, yielding attributions that satisfy the axioms of sensitivity and implementation invariance
76 (Sundararajan et al., 2017). Applying IG to the latent variables of a VAE trained on cancer expression
77 data identifies the top contributing genes of each latent dimension (Janizek et al., 2023; Qiu et al., 2025),
78 so that a latent dimension can be anchored to a pathway or cellular program. Combining a VAE with IG
79 therefore allows us to derive gene-level importance scores for the learned latent factors while retaining
80 interpretability.

81 1.5 Revealing a latent candidate regulator: the case of RAP1GAP2

82 Using this framework, our analysis identifies RAP1GAP2 as a gene with high latent-space attribution
83 that is highly predictive of tumour status, even though standard differential expression analysis does not
84 identify it as significant. RAP1GAP2 was first identified as a Rap1-specific GTPase-activating protein in
85 human platelets (Schultess et al., 2005). RAP1GAP2 has not been well studied in HNSCC, but it belongs
86 to the same family as RAP1GAP (RAP1GAP1), which has been reported to act as a tumour suppressor in
87 HNSCC: in oropharyngeal tumour models, restoring Rap1GAP expression reduced active Rap1
88 signalling and slowed tumour growth in vivo (Zhang et al., 2006). Conversely, Rap1GAP has also been
89 shown to promote invasion through MMP9 secretion in squamous cell carcinoma (Mitra et al., 2008).
90 This evidence of the Rap1 pathway's involvement in HNSCC provides context for our findings and
91 suggests that dysregulation of Rap1-inhibitory proteins could contribute to oncogenic processes.

92 1.6 Hypothesis

93 We hypothesised that deep neural networks, particularly VAEs, can learn latent representations of
94 transcriptomic data that capture complex, non-linear biological signals associated with HNSCC, and that
95 integrating unsupervised deep learning with Integrated Gradients can uncover gene-level contributors,
96 such as RAP1GAP2, that conventional DGE analysis overlooks.

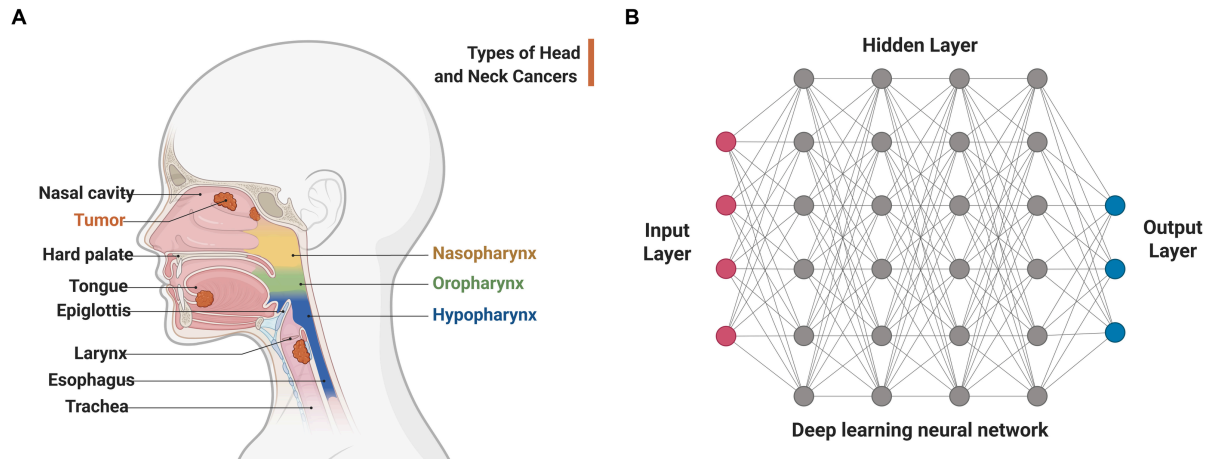
97 1.7 Aims and objectives

98 This study aimed to uncover hidden transcriptomic patterns and identify candidate gene-level regulators
99 of HNSCC by applying a DeepProfile-based VAE and Integrated Gradients framework. The objectives
100 were: (i) to compress gene expression into latent features with a VAE; (ii) to interpret the latent features
101 with Integrated Gradients; (iii) to characterise the latent features by pathway enrichment and GSEA; (iv)
102 to evaluate candidate genes with supervised deep learning classifiers; (v) to compare the results with
103 traditional DGE analysis; and (vi) to examine the top candidate, RAP1GAP2, across independent omic
104 layers including DNA methylation, copy number, mutation and survival.

105 2. Materials and methods

106 2.1 Workflow of the study

107 The study design is shown in Figure 2. Multi-dataset gene expression was transformed by PCA and
108 encoded by a VAE into a latent space, followed by Integrated Gradients-based gene attribution, pathway
109 analysis, supervised learning to identify candidate regulators, DGE analysis and multi-omic follow-up of
110 the top candidate.



111 **Figure 1.** Background. (A) Anatomical regions of the head and neck in which cancers arise, including the nasopharynx,
 112 oropharynx and hypopharynx. (B) Basic architecture of a deep neural network with input, hidden and output layers.
 113 [Created with Adobe Illustrator v27.8.1.]

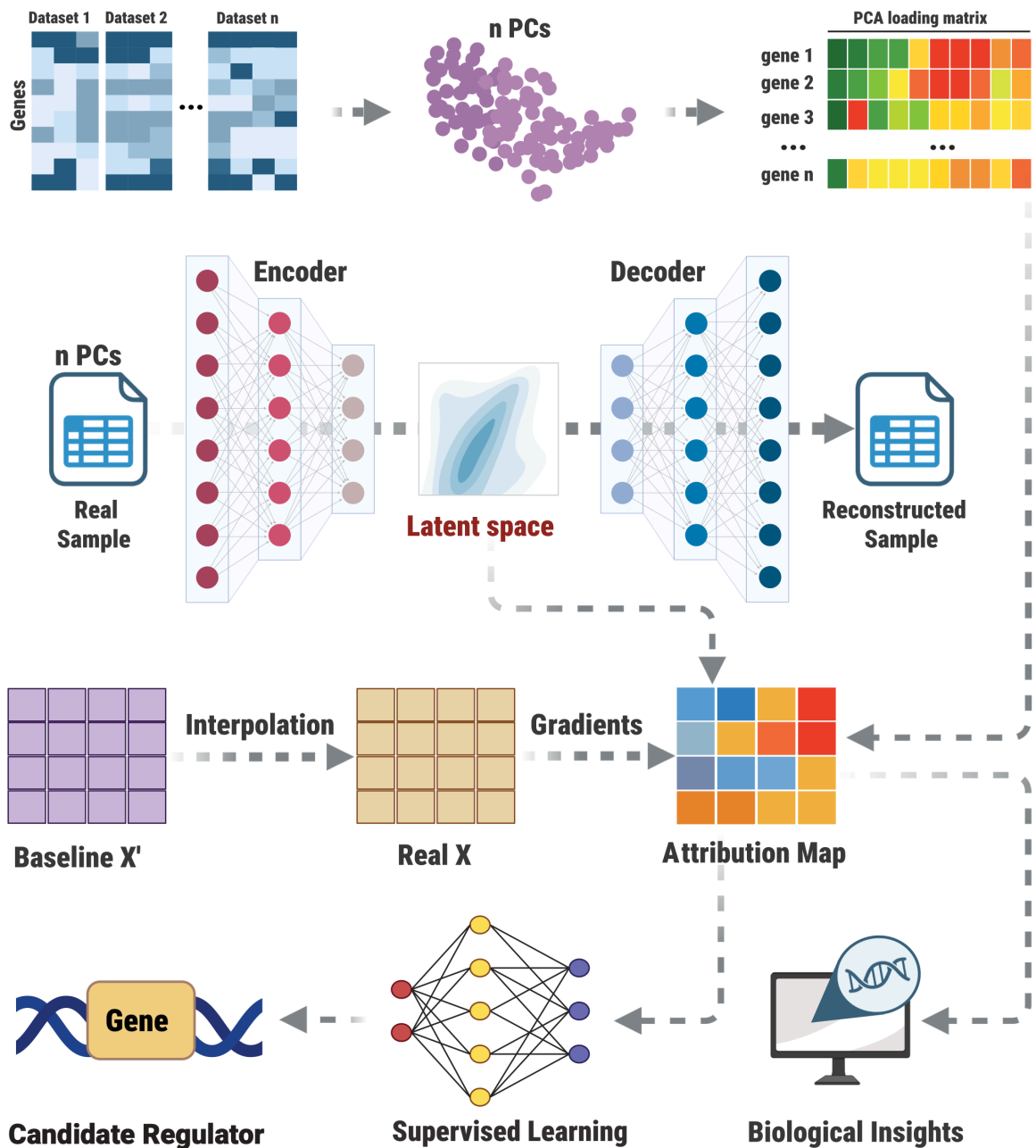


Figure 2. Workflow of the study. Multi-dataset gene expression is transformed by PCA and encoded by a VAE into a latent space; Integrated Gradients attributions over the latent space are mapped back to genes through the PCA loading matrix, followed by supervised learning to identify candidate regulators and extract biological insights. [Created with Adobe Illustrator v27.8.1.]

2.2 Datasets

We used the head and neck cancer expression compendium assembled and preprocessed for DeepProfile (Qiu et al., 2025). It comprises 643 samples from 26 publicly available GEO datasets generated on Affymetrix Human Genome U133 Plus 2.0 (GPL570), U133A (GPL96), U133B (GPL97), HG-Focus (GPL201) and Pathwork Diagnostics Pathchip (GPL7280) arrays, with 11,020 genes common to all datasets. The datasets are listed in Table 1; sample numbers in Table 1 are the totals of each GEO series, from which the head and neck cancer samples were included in the compendium.

125 **Table 1.** GEO datasets of the DeepProfile head and neck cancer compendium. Samples are the total number of samples in
126 each GEO series.

GEO accession	Samples in series	Platform	Study type
GSE100942	10	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE10300	44	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE12630	276	GPL7280 (Pathwork Diagnostics Pathchip); GPL96 (Affymetrix HG-U133A)	Expression profiling by array
GSE17351	10	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE2326	38	GPL97 (Affymetrix HG-U133B); GPL96 (Affymetrix HG-U133A)	Expression profiling by array
GSE27020	109	GPL96 (Affymetrix HG-U133A)	Expression profiling by array
GSE29330	18	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE30784	229	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE3292	36	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE33810	3	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE35603	78	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE42363	14	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE43580	150	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE46517	121	GPL96 (Affymetrix HG-U133A)	Expression profiling by array
GSE4975	14	GPL96 (Affymetrix HG-U133A); GPL97 (Affymetrix HG-U133B); GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE51010	56	GPL570 (Affymetrix HG-U133 Plus 2.0); GPL201 (Affymetrix HG-Focus)	Expression profiling by array
GSE51021	28	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE53355	44	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE5364	341	GPL96 (Affymetrix HG-U133A)	Expression profiling by array
GSE56330	2	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE60542	92	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE6791	84	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array

GSE74530	12	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE76039	37	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE78060	30	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array
GSE84713	28	GPL570 (Affymetrix HG-U133 Plus 2.0)	Expression profiling by array

2.3 Data integration and dimensionality reduction

Data integration, missing-value handling, batch correction and standardisation of the compendium were performed as described for DeepProfile (Qiu et al., 2025), resulting in expression values standardised per gene. Because of the high dimensionality of the expression matrix, principal component analysis (PCA; scikit-learn) was performed with 500 components, and both the transformed data (643 samples $\times$ 500 PCs) and the component loadings (11,020 genes $\times$ 500 PCs) were retained.

2.4 Variational autoencoder

A VAE was built on the 500-PC data to learn a compact, non-linear representation of gene expression. The encoder f maps each input x ($M = 500$ features) to the mean $\mu(x)$ and variance $\sigma(x)$ of a D -dimensional latent Gaussian ($D \ll M$), from which a latent vector $z \sim N(\mu(x), \sigma(x))$ is sampled; the decoder g reconstructs the input from z . The model minimises the squared Euclidean (L2) reconstruction error between x and $g(z)$, which measures the total reconstruction error across all input dimensions, plus the Kullback–Leibler divergence between $N(\mu(x), \sigma(x))$ and a standard normal prior, which regularises the latent space.

Following DeepProfile, the encoder had two fully connected hidden layers of 250 and 100 neurons, each followed by batch normalisation and ReLU activation; the decoder mirrored the encoder. Weights were initialised with the Glorot uniform method, and the model was trained with the Adam optimiser (learning rate 0.0005) in batches of 50 samples for 50 epochs (TensorFlow/Keras). The latent dimension was set to 50, and two independent models with different random seeds were trained. DeepProfile trains 100 models at each of six latent sizes and combines them into an ensemble; a single latent size and two runs were used here because of computational limits.

2.5 Gene attribution with Integrated Gradients

To determine which genes contribute to each latent variable, Integrated Gradients was applied to the encoder mean of every latent variable for every sample, with a zero baseline and 50 interpolation steps. Attributions over the 500 principal components were projected to gene level through the PCA loadings, and absolute gene attributions were averaged across samples, giving a gene $\times$ latent attribution matrix. Mean attribution across the 50 latent variables was used to rank genes globally.

2.6 Pathway enrichment and gene set enrichment analysis

To interpret the biological representation of each latent variable, top-ranked genes were analysed with g:Profiler through the gprofiler2 R package (Peterson et al., 2020), mapping genes to Gene Ontology terms, KEGG pathways and Reactome pathways (Homo sapiens; Benjamini–Hochberg FDR < 0.05 ; electronic GO annotations excluded). Gene set enrichment analysis (GSEA) was performed with pre-ranked gene lists derived from the attributions of each latent variable against KEGG pathways

(Subramanian et al., 2005). The normalised enrichment score (NES) of each pathway–latent pair was assembled into a matrix, and the 50 pathways with the highest variance across latent variables were visualised as a heatmap (seaborn).

2.7 Supervised deep learning classification

The 20 genes with the highest mean attribution were selected as candidate regulators and evaluated in independent HNSCC data comprising GSE178537 (bulk RNA-seq of primary tumour, metastatic and adjacent normal tissue; Illumina NextSeq 500 and HiSeq 2500) and GSE290057 (GeoMx whole-transcriptome spatial profiling; Illumina NovaSeq 6000). Gene-wise z-score normalisation was applied within each dataset, followed by empirical Bayes batch correction (ComBat; Johnson et al., 2007) implemented in pyComBat (Behdenna et al., 2023), using pandas and numpy. Samples were labelled as tumour (HNSCC) or control.

Three deep learning models were compared with stratified five-fold cross-validation (TensorFlow/Keras): a multilayer perceptron (MLP; 64 and 32 ReLU units, dropout 0.3), a one-dimensional convolutional neural network (CNN; 32 filters, kernel size 3, max pooling, 32-unit dense layer) and a long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997; 32 units), each with a sigmoid output, binary cross-entropy loss, Adam optimiser (learning rate 0.001), batch size 16, up to 50 epochs and early stopping on validation loss (patience 5; 20% validation split). Gene importance was estimated by permutation importance (AUROC; 30 repeats) for a grid-searched MLP classifier (scikit-learn). Single-gene MLP classifiers were trained for each gene (grid search over 32 or 64 units, dropout 0.2 or 0.3, batch size 8 or 16, 30 epochs). Performance was evaluated as the area under the precision–recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC), averaged across folds, and visualised with matplotlib and seaborn.

2.8 Differential gene expression analysis

RNA-seq count data were analysed with DESeq2 (Love et al., 2014). Genes with low counts were removed before normalisation, log2 fold changes were shrunk with apleglm (Zhu et al., 2019), and significance was defined as adjusted $p < 0.05$ and $|\log_2 \text{fold change}| > 1$. Volcano plots were generated with ggplot2.

2.9 RAP1GAP2 multi-omic analysis

Somatic mutations, GISTIC copy number, RNA-seq expression, promoter methylation (HM450) and clinical data for TCGA-HNSC (Cancer Genome Atlas Network, 2015) were retrieved from cBioPortal (Cerami et al., 2012; studies hnscc_tcga_pan_can_atlas_2018 and hnscc_tcga). Tumour-versus-normal differential expression was computed from GDC STAR gene counts for the 43 TCGA-HNSC patients with a primary tumour and a matched adjacent normal sample, and replicated in GSE178537 (RSEM expected counts; 20 patients with paired tumour and adjacent normal tissue; lymph-node metastases excluded), using PyDESeq2 (Muzellec et al., 2023) with the design ~patient + condition and retaining genes with at least 10 counts in at least 10 samples. Within tumours, RAP1GAP2 expression tertiles were compared for five downstream markers (CCND1, CDK4, CDK6, MAPK1, MAPK3) with Mann–Whitney U tests and Holm correction. Overall survival was compared between expression tertiles with Kaplan–Meier curves, the log-rank test and Cox regression (lifelines).

2.10 Genome-wide promoter methylation and integration with expression

TCGA-HNSC Illumina HumanMethylation450 (Bibikova et al., 2011) β values (528 primary tumours, 50 solid-tissue normals), STAR gene counts and clinical data were downloaded from the GDC with TCGAAbiolinks (Colaprico et al., 2016). Non-CpG probes, probes on sex chromosomes, probes overlapping a common SNP ($MAF \geq 0.05$) at the CpG or single-base-extension site, and probes missing in more than 10% of samples were removed (394,533 of 485,577 retained). Differential methylation was tested with limma on M-values (Ritchie et al., 2015); CpGs with $FDR < 0.05$ and $|\Delta\beta| \geq 0.2$ were called hyper- or hypomethylated. CpGs annotated (minfi; Aryee et al., 2014) to TSS1500, TSS200, 5'UTR or the first exon were treated as promoter CpGs, and a gene was called promoter-hyper- or hypomethylated when at least two promoter CpGs were significant in the same direction and none in the opposite direction. DEGs were computed with DESeq2 ($FDR < 0.05$, $|\log_2FC| \geq 1$). For overlapping genes, Spearman correlation between mean promoter β and $\log_2(CPM + 1)$ was computed in the 540 samples with both data types (Benjamini–Hochberg correction), and overlap enrichment was tested with one-sided Fisher's exact tests among the 14,066 genes measured on both platforms.

2.11 Code availability

Analyses were performed in Python (TensorFlow/Keras, scikit-learn, pandas, PyDESeq2, lifelines) and R (DESeq2, limma, minfi, TCGAAbiolinks, gprofiler2, ggplot2). All code is available at <https://github.com/Prokash21/Deep-Neural-Profiling>.

3. Results

3.1 Data preprocessing and quality assessment

Highly expressive models such as deep neural networks tend to overfit when the sample size is small; we therefore used a compendium of 26 head and neck cancer expression datasets from different platforms. Standardised expression values of the 11,020 genes showed consistent distributions across the 643 samples (Figure 3A). The scatterplot of the first two principal components showed no clear clustering or batch structure (Figure 3B), indicating that the data were suitable for unsupervised modelling. The 500 principal components were used as input to the VAE.

3.2 Latent space extraction using a deep neural network

VAE models with 50 latent variables were trained, and Integrated Gradients attributions were computed for every latent variable. As an example, Figure 3C shows the 10 genes with the highest attribution to latent variable 0: MSRB1, TTI2, MPP5, ATF3, PYHIN1, HPCAL1, PICALM, COMMD8, SFT2D2 and RNF130, with attribution scores of approximately 0.0041–0.0042. These genes are the primary contributors to the signal captured by this latent variable.

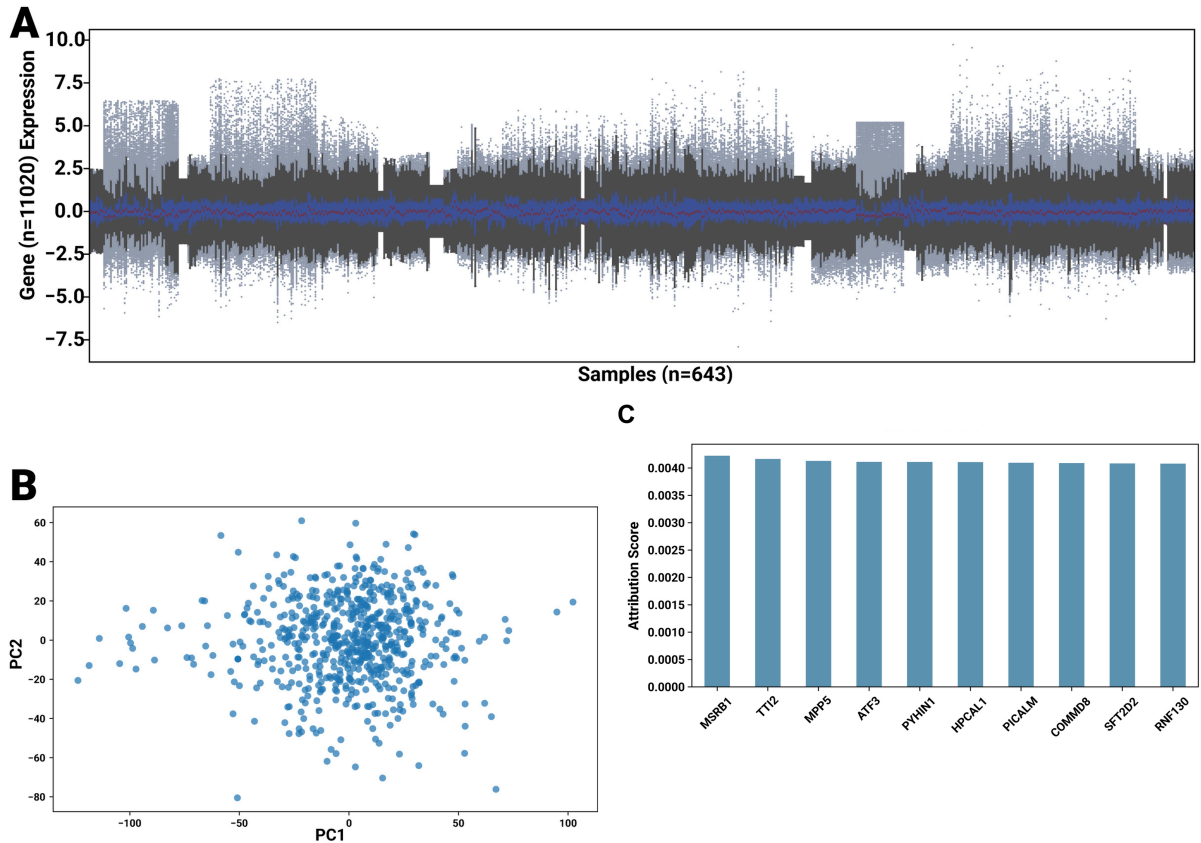


Figure 3. Data quality and latent attribution. (A) Boxplot of standardised expression values of 11,020 genes across the 643 samples; each box is one sample and dots are outliers. (B) Scatterplot of the first two principal components. (C) The 10 genes with the highest Integrated Gradients attribution to latent variable 0.

3.3 Latent variables capture distinct gene programs and biological pathways

To characterise the biological meaning of the latent space, we computed the mean attribution of each gene across all 50 latent variables and selected the 20 genes with the highest overall contribution: DDX43, FABP4, RP11, RAP1GAP2, WDR41, KCNK5, XIST, SAA2-SAA4, ZNF839, CTH, ERC2, GATA3, KIAA0020, PDK3, DTWD1, RPARP-AS1, ENO3, HAPLN1, GSTT2 and SLC9A7 (Figure 4A). Mean attribution scores ranged from approximately 0.0035 to 0.0055. Hierarchical clustering of latent variables by these attribution profiles revealed modular structures in which sets of genes contributed jointly to subsets of latent variables. The highest attributions were concentrated in a few latent variables: for example, latent 14 (CTH, RPARP-AS1), latent 26 (CTH, GATA3, PDK3) and latent 9, which had the strongest attribution from RAP1GAP2.

g:Profiler enrichment of the top-ranked genes of individual latent variables linked several latent variables to distinct pathways (Figure 4B; Table 3). Latent 9 was enriched for DNA integration, suggesting a link to genomic stability or viral interaction processes. Latent 20 was enriched for Golgi lumen acidification and *Vibrio cholerae* infection, indicating a role in intracellular trafficking. Latent 7 was associated with regulation of nervous system development, latent 31 with ECM–receptor interaction, pointing toward microenvironmental and adhesion-related mechanisms, and latents 42 and 43 with base excision repair and RNA degradation, respectively. The strongest enrichment was cellular response to 2,3,7,8-tetrachlorodibenzodioxine in latent 41 (adjusted $p = 0.0038$).

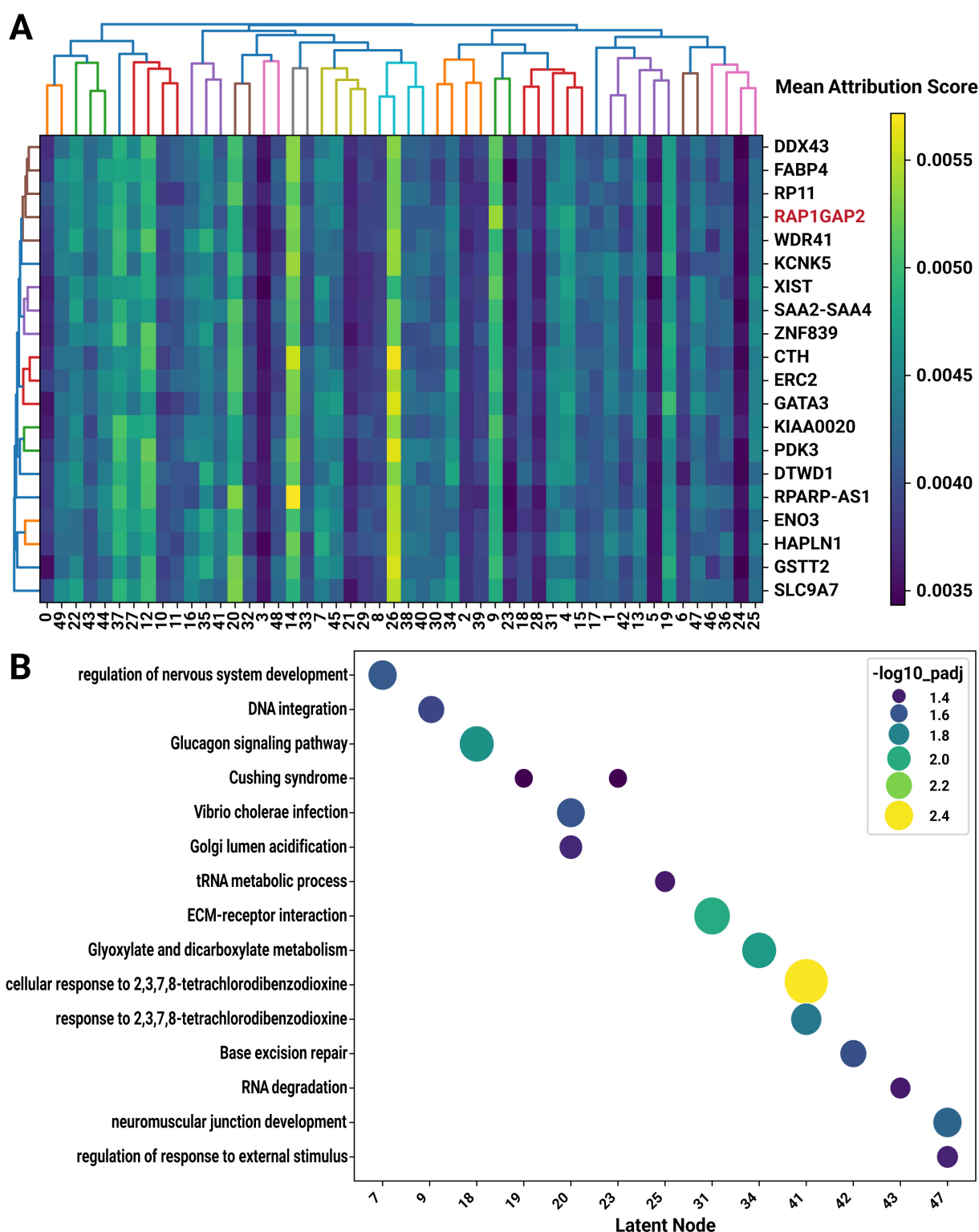


Figure 4. Interpretation of VAE latent variables through gene attribution and pathway enrichment. (A) Heatmap of mean Integrated Gradients attribution scores of the 20 top-ranked genes across all 50 latent variables; rows (genes) and columns (latent variables) are hierarchically clustered, and RAP1GAP2 is highlighted. (B) g:Profiler enrichment of top-ranked genes for selected latent variables; dot size and colour show $-\log_{10}$ adjusted p.

256 **Table 3.** Selected g:Profiler enrichments of latent variables (adjusted $p < 0.05$).

Latent	Term	Adjusted p	Term size	Query size	Intersection
7	regulation of nervous system development	0.0236	457	16	5
9	DNA integration	0.0288	10	15	2
18	Glucagon signaling pathway	0.0135	107	10	3
19	Cushing syndrome	0.0500	153	11	3
20	Vibrio cholerae infection	0.0247	50	6	2
20	Golgi lumen acidification	0.0371	13	13	2
23	Cushing syndrome	0.0500	153	11	3
25	tRNA metabolic process	0.0414	210	18	4
31	ECM-receptor interaction	0.0101	89	11	3
34	Glyoxylate and dicarboxylate metabolism	0.0122	30	7	2
41	cellular response to 2,3,7,8-tetrachlorodibenzodioxine	0.0038	4	15	2
41	response to 2,3,7,8-tetrachlorodibenzodioxine	0.0179	8	15	2
42	Base excision repair	0.0264	44	7	2
43	RNA degradation	0.0412	79	5	2
47	neuromuscular junction development	0.0215	54	19	3
47	regulation of response to external stimulus	0.0391	1082	19	7

257 **3.4 Functional characterisation of latent variables by GSEA**

258 GSEA of the ranked gene attributions of each latent variable is summarised as an NES heatmap of the 50
259 most variable KEGG pathways (Figure 5). Latent 36 showed the strongest positive enrichment, for Cell
260 cycle and DNA replication ($NES \approx 3$), together with positive enrichment of Fanconi anemia pathway and
261 Mismatch repair, indicating a proliferation and genome-maintenance program. Ribosome showed
262 positive enrichment in latent 21 and strong negative enrichment in latents 6 and 29–33, while Proteasome
263 was negatively enriched in most latent variables. Herpes simplex virus 1 infection was positively
264 enriched across many latent variables, and Non-alcoholic fatty liver disease (latent 25) and Cardiac
265 muscle contraction (latent 28) showed isolated positive enrichment. Immune-related pathways such as
266 JAK-STAT signalling and Cytokine–cytokine receptor interaction showed weaker, mixed enrichment.

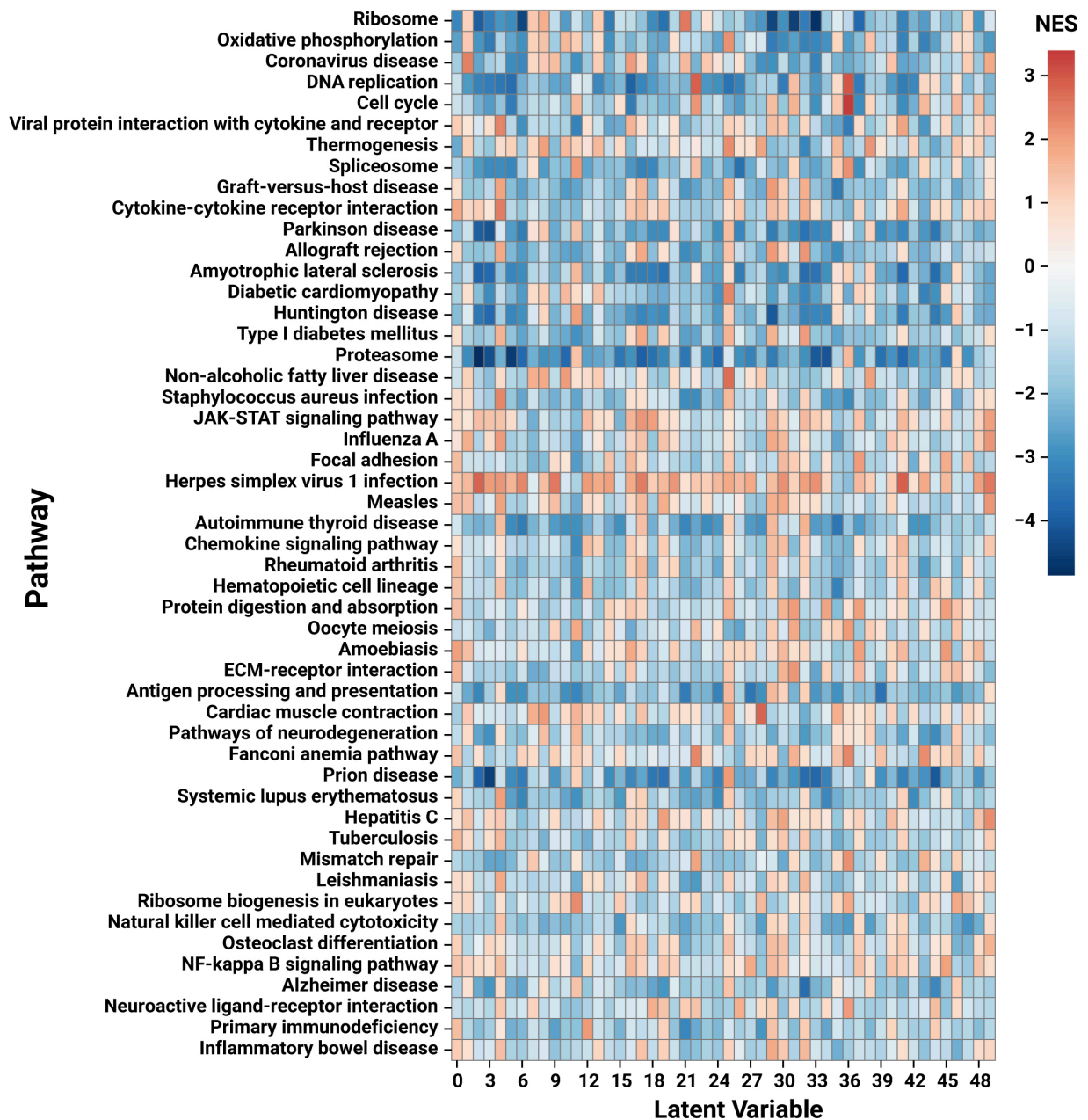


Figure 5. Pathway enrichment heatmap of VAE latent variables using GSEA. Normalised enrichment scores (NES) of the 50 most variable KEGG pathways (rows) across the 50 latent variables (columns); red indicates positive and blue negative enrichment.

3.5 Deep learning-based classification of candidate genes

Eighteen of the 20 candidate genes were measured in the independent validation data (RP11 and KIAA0020 were absent). Their expression distributions in control and tumour samples are shown as violin plots in Figure 6A; most genes showed overlapping distributions between the two groups. We then assessed the ability of the 18-gene panel to classify tumour samples with supervised deep learning models (Figure 6B, C). The MLP performed best, with a mean AUPRC of 0.86 and mean AUROC of 0.80, followed by the CNN (AUPRC 0.81, AUROC 0.79) and the LSTM (AUPRC 0.78, AUROC 0.70).

3.6 RAP1GAP2 is the most predictive gene in single-feature classification

To identify the most predictive gene in the panel, we computed permutation feature importance for the MLP model (Figure 6D). RAP1GAP2 ranked first (importance ≈ 0.050), followed closely by XIST and

SLC9A7, and then FABP4 and KCNK5, indicating that RAP1GAP2 carries the strongest discriminative information for separating tumour from control samples. A single-gene MLP classifier trained only on RAP1GAP2 expression achieved a mean AUPRC of 0.769 (Figure 6E). Performance metrics of the single-gene models are listed in Table 2; RAP1GAP2 had the highest AUROC (0.710, shared with SLC9A7), AUPRC (0.769) and accuracy (0.730).

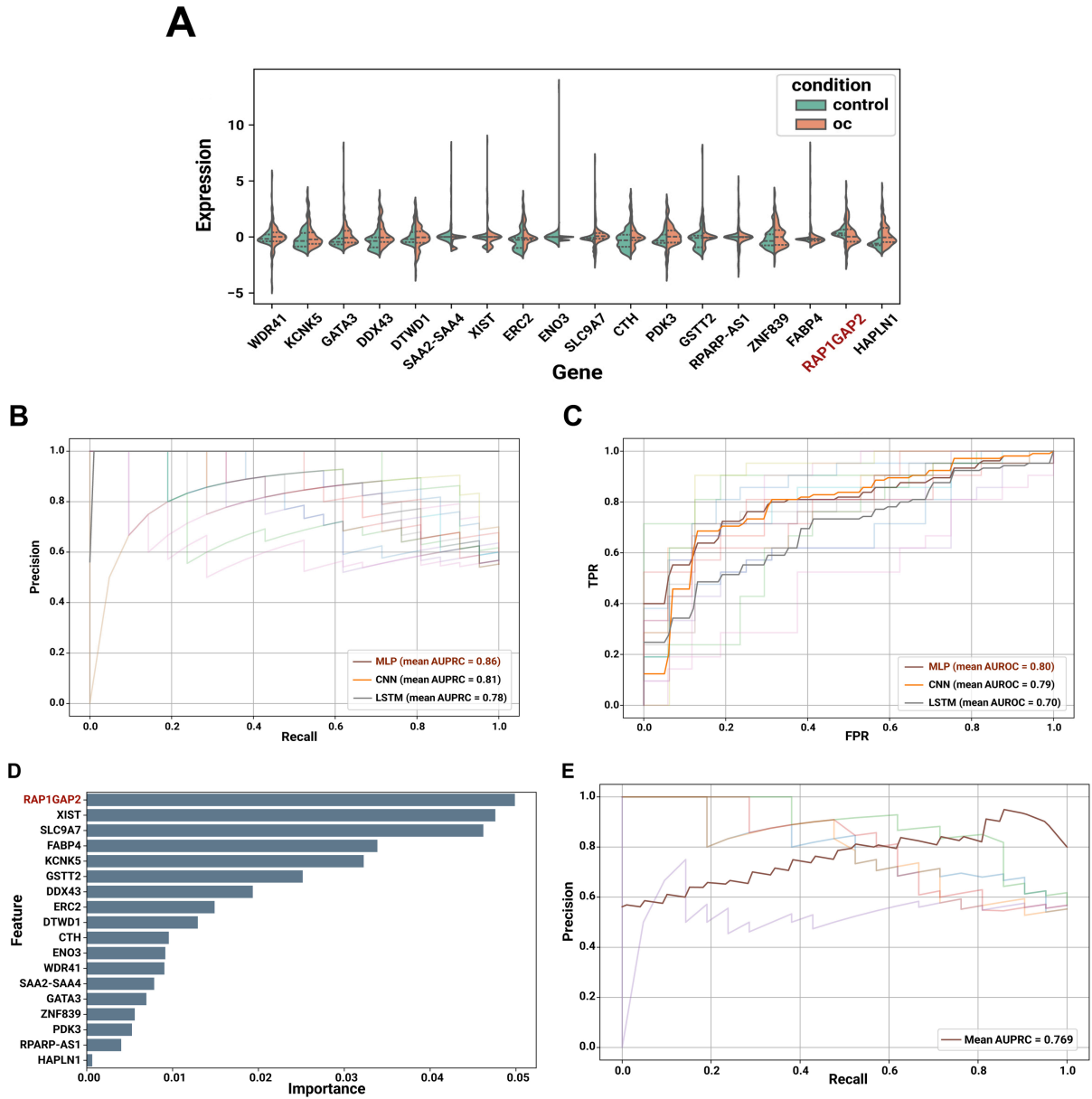


Figure 6. Classification of tumour and control samples with the candidate genes in the independent validation data. (A) Violin plots of the expression of the 18 measured candidate genes in control and tumour (HNSCC; labelled “oc”) samples. (B) Precision–recall and (C) ROC curves of MLP, CNN and LSTM models trained on the 18 genes (faint lines, individual folds). (D) Permutation feature importance of each gene in the MLP model. (E) Precision–recall curves of the single-gene classifier trained on RAP1GAP2 expression (mean AUPRC 0.769).

290 **Table 2.** Performance metrics of single-gene classification models.

Gene	AUROC	AUPRC	Accuracy	F1	Precision	Recall
WDR41	0.640	0.650	0.556	0.711	0.560	0.971
KCNK5	0.540	0.560	0.524	0.686	0.545	0.924
GATA3	0.590	0.580	0.620	0.667	0.657	0.676
DDX43	0.550	0.570	0.513	0.629	0.550	0.733
DTWD1	0.650	0.640	0.535	0.679	0.554	0.876
XIST	0.594	0.708	0.540	0.688	0.556	0.905
SAA2-SAA4	0.557	0.621	0.556	0.709	0.561	0.962
ERC2	0.550	0.560	0.610	0.709	0.610	0.848
ENO3	0.610	0.590	0.567	0.722	0.565	1.000
SLC9A7	0.710	0.710	0.594	0.689	0.604	0.800
CTH	0.590	0.590	0.604	0.711	0.603	0.867
PDK3	0.570	0.600	0.567	0.675	0.583	0.800
GSTT2	0.643	0.664	0.642	0.735	0.628	0.886
RPARP-AS1	0.528	0.619	0.540	0.699	0.552	0.952
ZNF839	0.550	0.590	0.556	0.711	0.560	0.971
FABP4	0.520	0.530	0.535	0.695	0.550	0.943
RAP1GAP2	0.710	0.769	0.730	0.760	0.700	0.940
HAPLN1	0.606	0.676	0.615	0.692	0.628	0.771

291 **3.7 RAP1GAP2 is not detected by differential expression analysis**

292 Figure 7A and B show expression distributions of the RNA-seq data before and after DESeq2
293 normalisation, which made the distributions comparable across samples. DGE analysis identified 359
294 upregulated and 238 downregulated genes (Figure 7C). RAP1GAP2 did not meet the fold-change
295 threshold ($|\log_2FC| > 1$) and was therefore not called differentially expressed, although it was the top
296 contributor in the classification models.

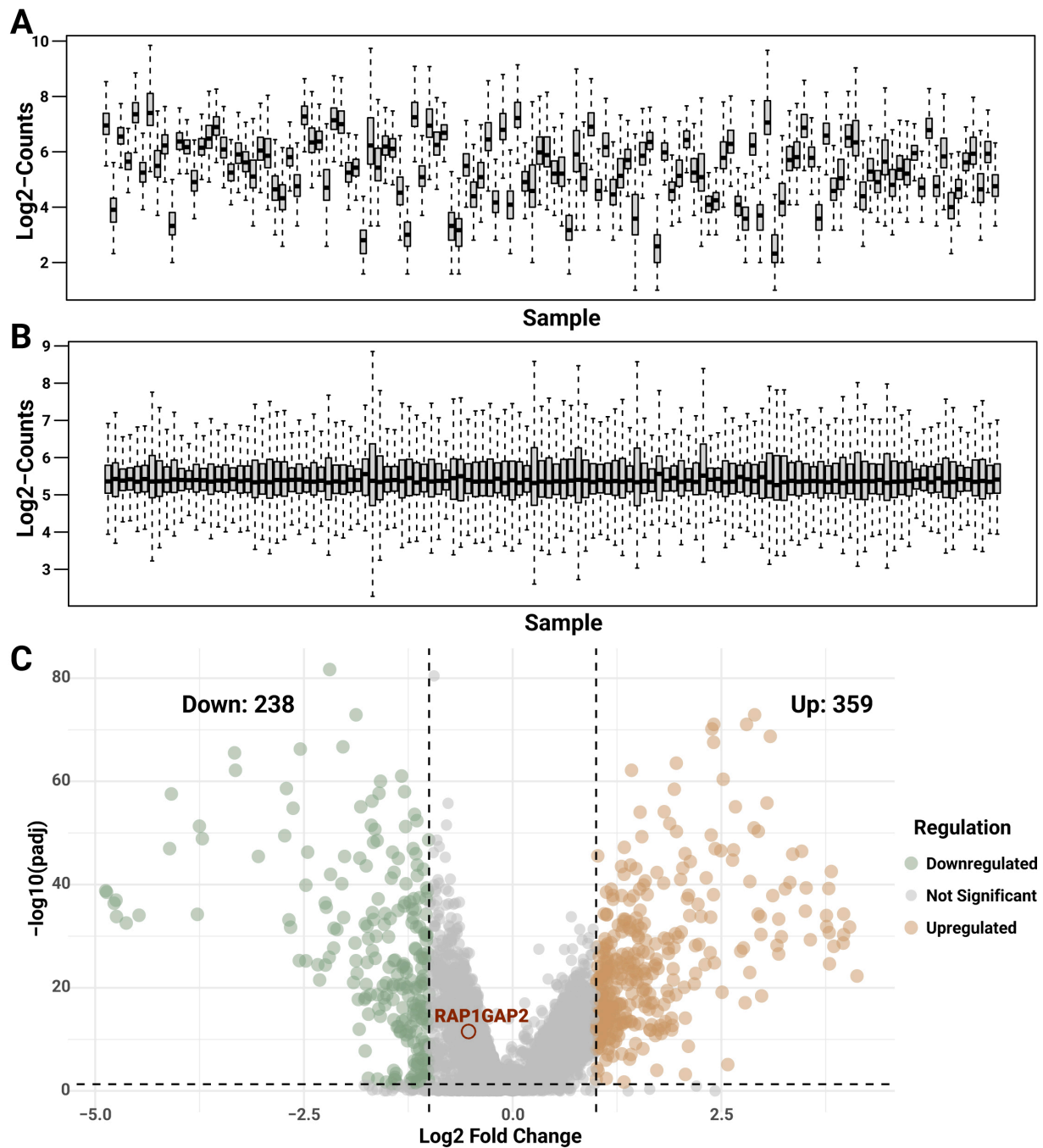


Figure 7. RAP1GAP2 is not identified by differential expression analysis. (A) Log2 counts per sample before normalisation. (B) Log2 counts after DESeq2 normalisation. (C) Volcano plot of differential expression (adjusted $p < 0.05$, $|\log_2\text{FC}| > 1$; 359 up, 238 down); RAP1GAP2 is circled and falls within the non-significant region.

3.8 RAP1GAP2 is unchanged in paired tumour–normal tissue in two independent cohorts

To test the DGE result in independent paired data, we compared tumours with matched adjacent normal tissue. In 43 TCGA-HNSC patients, RAP1GAP2 expression was unchanged ($\log_2\text{FC} -0.13$, 95% CI -0.49 to 0.24 ; adjusted $p = 0.57$), as was that of its paralog RAP1GAP ($\log_2\text{FC} -0.08$; adjusted $p = 0.79$). The result replicated in 20 paired GSE178537 patients (RAP1GAP2 $\log_2\text{FC} +0.31$, 95% CI -0.08 to 0.70 , adjusted $p = 0.17$; RAP1GAP $\log_2\text{FC} +0.01$, adjusted $p = 0.99$) (Figure 8A). In the same TCGA model, established HNSCC markers (MMP1 and CDKN2A up; KRT4 and CRNN down) reached adjusted $p < 10^{-20}$, confirming that the analysis detects real tumour–normal differences.

3.9 RAP1GAP2 promoter methylation, copy number, mutation and survival

In the GDC 450K data (528 tumours, 50 normals), 7 of the 10 RAP1GAP2 promoter CpGs were significantly hypomethylated in tumours (mean $\Delta\beta$ -0.20 ; minimum FDR 1.3×10^{-47}), whereas the RAP1GAP promoter showed no CpG passing $|\Delta\beta| \geq 0.2$. Despite this, RAP1GAP2 expression was similar in tumours and normals of the same samples (Figure 8B), and mean promoter β was not correlated with expression across tumours and normals combined (Spearman $\rho = 0.01$; 540 samples). Within tumours, promoter methylation was weakly negatively associated with expression ($\rho = -0.26$, $p = 2.1 \times 10^{-9}$, $n = 522$; Figure 8C).

RAP1GAP2 was mutated in 1 of 515 sequenced tumours (0.19%; missense V335A within the Rap/Ran-GAP domain), compared with 6 of 515 (1.17%) for RAP1GAP. Among 517 tumours with copy-number calls, RAP1GAP2 had 66 single-copy gains (12.8%) and 121 single-copy losses (23.4%), with no high-level amplification and one deep deletion. RAP1GAP2 expression tertiles did not stratify overall survival (log-rank $p = 0.91$; Cox HR low vs high 1.02, 95% CI 0.74–1.40; Figure 8D). Within tumours, high RAP1GAP2 expression was associated with higher CDK6 (Holm-adjusted $p = 9.2 \times 10^{-5}$) and MAPK1 ($p = 9.3 \times 10^{-4}$) and lower MAPK3 ($p = 2.2 \times 10^{-3}$) expression, whereas CCND1 and CDK4 did not differ after correction.

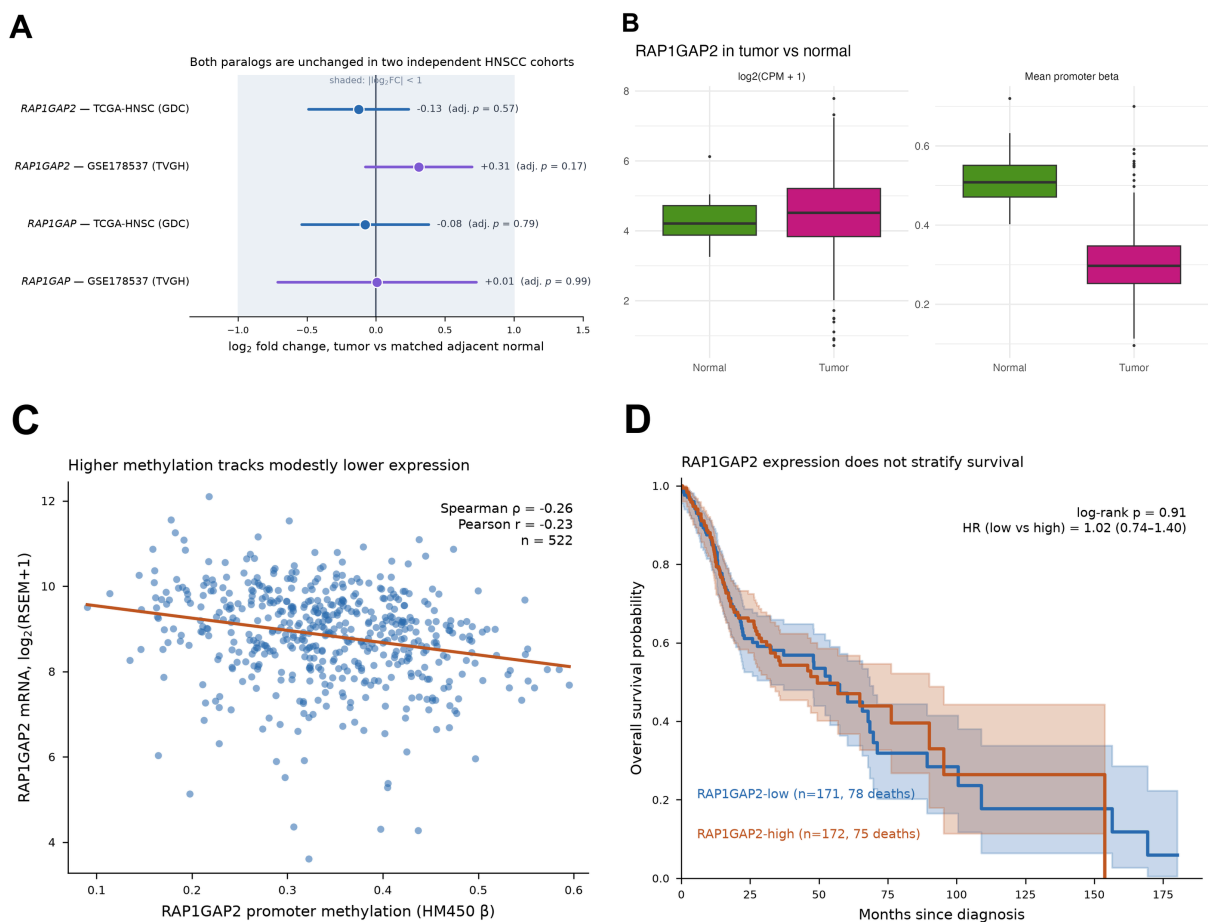


Figure 8. Multi-omic analysis of RAP1GAP2. (A) Tumour versus matched adjacent normal $\log_2$ fold changes (95% CI) of RAP1GAP2 and RAP1GAP in TCGA-HNSC (43 pairs) and GSE178537 (20 pairs); the shaded band marks $|\log_2FC| < 1$. (B) RAP1GAP2 expression ($\log_2$ CPM + 1) and mean promoter β value in the 540 TCGA-HNSC samples with both data types (20 normal, 520 tumour). (C) RAP1GAP2 promoter methylation versus mRNA expression in TCGA-HNSC tumours ($n = 522$). (D) Kaplan–Meier overall survival for low versus high RAP1GAP2 expression tertiles.

3.10 Genome-wide integration of promoter methylation and expression

Across TCGA-HNSC, 14,603 CpGs were hypermethylated and 14,771 hypomethylated in tumours; at the gene level, 999 genes had hypermethylated and 772 hypomethylated promoters (Supplementary Figure S1A). DESeq2 identified 4,182 upregulated and 2,759 downregulated genes. Overlaps were larger than expected by chance for hypermethylated-downregulated genes (187 observed vs 95.9 expected; odds ratio 2.41; Fisher $p = 9.2 \times 10^{-21}$) and hypomethylated-upregulated genes (91 vs 39.9; odds ratio 2.87; $p = 1.1 \times 10^{-14}$) (Supplementary Figure S1B–D). The discordant hypermethylated-upregulated overlap was also enriched (155 vs 90.3; odds ratio 2.00; $p = 3.2 \times 10^{-12}$). Requiring a significant negative methylation–expression correlation left 163 hypermethylated-downregulated genes (for example EPHX3, ZNF471, ZNF568, SVIP and SELENBP1) and 42 hypomethylated-upregulated genes (for example BST2, AIM2, GTSF1 and MMP13) (Supplementary Table S5).

4. Discussion

This study employed a deep learning framework to reveal molecular patterns and latent candidate regulators that traditional methods overlook in head and neck cancer transcriptomes. Training a VAE on a 643-sample compendium, we reduced the data to 50 latent variables that preserved essential variation among samples. Integrated Gradients quantified the contribution of each gene to each latent variable and revealed high-attribution gene sets that were enriched for biological pathways (Figures 4 and 5). Several latent variables were associated with processes such as cell adhesion, extracellular matrix interaction, DNA repair and cell-cycle progression, mechanisms commonly implicated in tumour progression, and high-contribution genes recurred across multiple latent variables, indicating that shared biological programs are embedded within distinct transcriptomic patterns.

RAP1GAP2 emerged as a consistent and discriminative component among the genes contributing to these latent traits (Figure 6). Despite its high latent-space attribution and its performance as a single-gene classifier (AUPRC 0.769), RAP1GAP2 was not significant in the differential expression analysis (Figure 7C), and paired analyses in two independent cohorts confirmed that its expression level does not change between tumour and normal tissue (Figure 8A). This disparity between statistical insignificance and predictive importance underscores the value of the approach: deep generative models can highlight genes whose contribution is not captured by average expression differences.

Our results relate to existing knowledge of Rap1 signalling. The Ras-related GTPase Rap1 and its regulators influence adhesion, secretion and motility (Bos et al., 2001), and active Rap1 signalling has been shown to enhance the invasiveness of head and neck malignancies through β -catenin and MMP7 (Zhang et al., 2017). Conversely, the Rap1 inactivator RAP1GAP, a paralog of RAP1GAP2, inhibits Rap1–ERK signalling and tumour proliferation in HNSCC models derived from the oropharynx (Zhang et al., 2006), whereas in other contexts elevated Rap1GAP enhances invasiveness through MMP9 secretion (Mittra et al., 2008). Rap1 regulators can therefore exert opposite effects in different cellular contexts (Zhang et al., 2017). RAP1GAP2 has not previously been examined in HNSCC, and our latent-space profiling identifies it as a candidate for study.

We also identified other latent-ranked genes, including PDK3 and FABP4, that are consistent with known cancer biology. PDK3 (pyruvate dehydrogenase kinase 3) mediates the Warburg effect and is increased in hypoxic malignancies, contributing to metabolic reprogramming and aggressive behaviour (Lu et al., 2011). FABP4 (fatty acid-binding protein 4) facilitates tumour metastasis and treatment resistance by enhancing lipid transport and signalling (Sun and Zhao, 2022). The presence of these genes among the

top latent contributors suggests that the model captured characteristics of cancer such as metabolic plasticity and microenvironmental adaptation alongside the Rap1 signalling axis.

Based on the known biochemistry of RAP1GAP2, we propose a model for how it could contribute to tumour invasion (Figure 9). RAP1GAP2 is a GTPase-activating protein for Rap1 that converts active GTP-bound Rap1 into its inactive GDP-bound state. Active Rap1 stabilises integrin- and E-cadherin-mediated adhesion (Price et al., 2004), so its inactivation could weaken cell–cell adhesion. Loss of Rap1 activity could also relieve Rap1-mediated restraint of Ras–MAPK/ERK signalling, which promotes proliferation, migration and expression of matrix metalloproteinases (Zhang et al., 2017; Mitra et al., 2008). In addition, RAP1GAP2 interacts with synaptotagmin-like protein 1 (Slp1) and Rab27 to regulate secretory vesicles (Neumüller et al., 2009; Li et al., 2018), which could influence the release of matrix-degrading enzymes such as MMP-2 and MMP-9. This model is a hypothesis derived from the literature; it was not tested in this study.

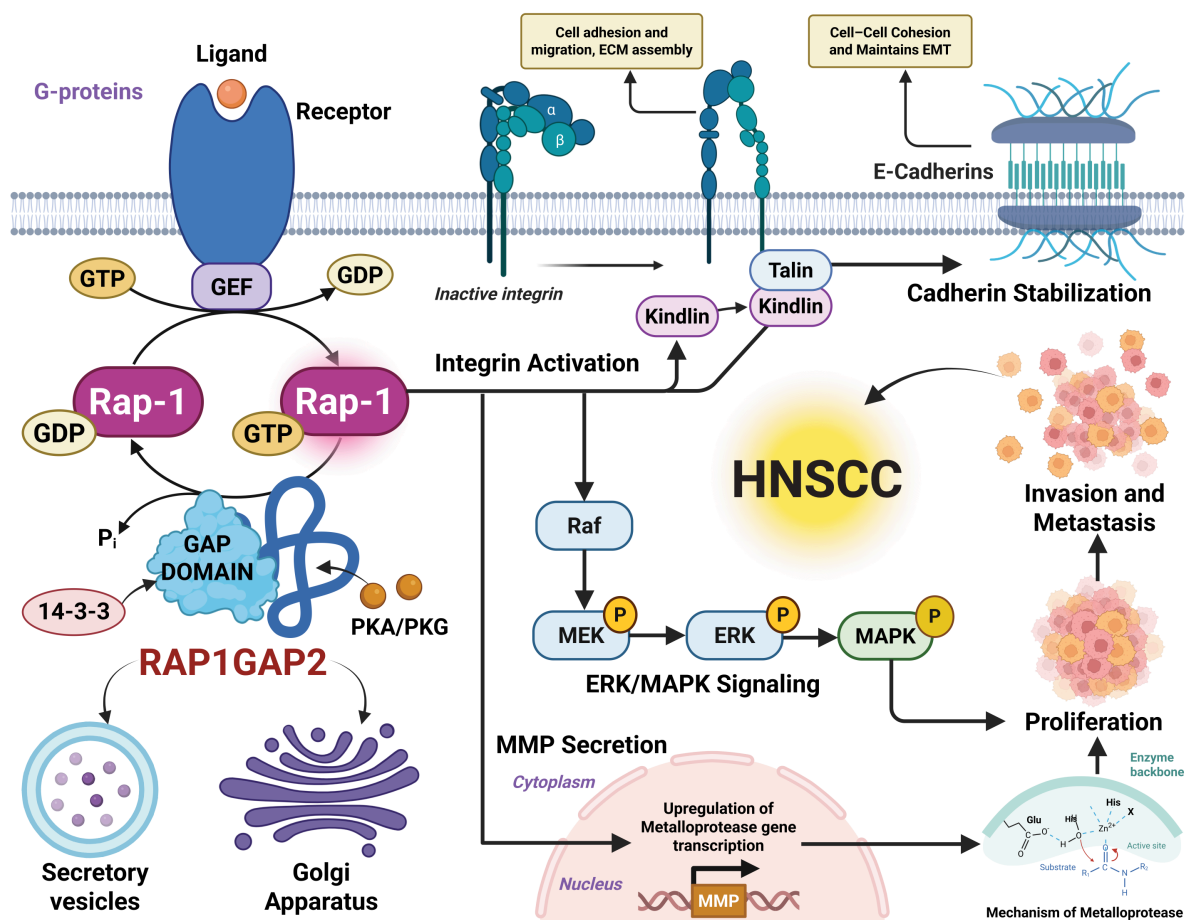


Figure 9. Proposed model of RAP1GAP2 function in HNSCC invasion, based on published biochemistry. RAP1GAP2 inactivates Rap1, which may weaken integrin- and cadherin-mediated adhesion, relieve restraint of ERK/MAPK signalling and, through Golgi-derived secretory vesicles, promote MMP secretion. The model is hypothetical and was not tested in this study. [Created with Adobe Illustrator v27.8.1.]

The multi-omic follow-up refines this picture. RAP1GAP2 is almost never mutated in TCGA-HNSC, its copy-number changes are shallow, and its expression is not associated with overall survival. Its promoter is clearly hypomethylated in tumours, but this change is not accompanied by a change in expression, and the within-tumour association between methylation and expression is weak. Within tumours, RAP1GAP2 expression tracked CDK6 and MAPK1 positively and MAPK3 negatively; because MAPK1 and

MAPK3 encode ERK2 and ERK1 and moved in opposite directions, these transcript-level associations do not establish an effect on ERK pathway activity. Together, these results indicate that the contribution of RAP1GAP2 to the latent space is not explained by genetic or promoter-methylation changes and may instead reflect its co-variation with other genes, cell-type composition, or regulation at the protein level, which requires experimental investigation. The genome-wide analysis additionally provides 205 genes with concordant promoter methylation and expression changes and a significant negative correlation, a resource for future studies of epigenetic regulation in HNSCC.

4.1 Limitations

Several limitations should be noted. First, functional experiments were not performed, so the role of RAP1GAP2 in invasion and metastasis remains to be demonstrated; our results are correlative. Second, the VAE was trained with two runs at a single latent size of 50 rather than the full DeepProfile ensemble, and the latent dimensionality was not optimised; attribution rankings may change with more runs. Third, the discovery compendium integrates 26 microarray datasets from different platforms and head and neck subsites; tumour subsite, treatment history and HPV status were not modelled and could affect the extracted latent features. Fourth, the validation data combine bulk RNA-seq and spatial transcriptomic profiles, and classification performance should be confirmed in a larger, homogeneous cohort. Fifth, candidate selection prioritised predictive power over statistical significance, which may introduce false positives. Sixth, adjacent normal mucosa in HNSCC can carry field-cancerisation changes, biasing tumour–normal comparisons toward the null. Finally, post-translational regulation of RAP1GAP2 was not examined.

4.2 Future directions

Functional validation is essential. Knockdown or overexpression of RAP1GAP2 in HNSCC cell lines, followed by measurement of invasion, Rap1-GTP activity and downstream signalling (ERK/MAPK activation and MMP secretion), would directly test the proposed model, and in vivo models could further assess a pro-metastatic role. Validation of RAP1GAP2 expression by qRT-PCR in HNSCC specimens is under way. Assessing RAP1GAP2 at the protein level in larger cohorts or tissue arrays could reveal associations with tumour stage, metastasis or HPV status. Methodologically, the deep profiling framework can be extended with the full DeepProfile ensemble, applied to HPV-stratified datasets, or integrated with proteomic and epigenetic data to find cross-layer or context-specific candidate regulators, and the interactions between latent-ranked genes such as RAP1GAP2, PDK3 and FABP4 merit systems-level study.

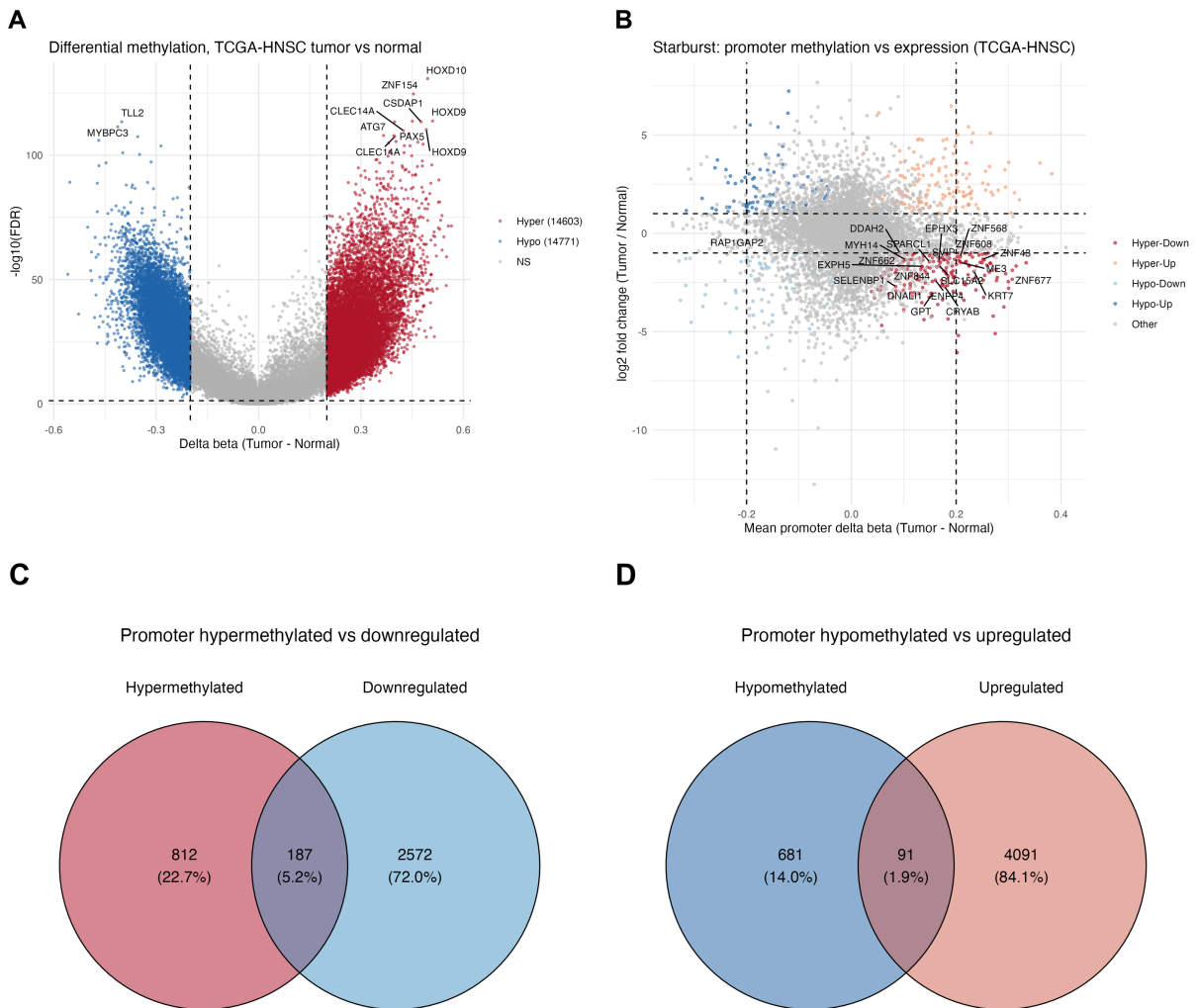
5. Conclusion

This study demonstrates that combining deep learning with high-dimensional transcriptome data can prioritise genes that conventional analysis overlooks. RAP1GAP2, a gene without differential expression between tumour and normal tissue, emerged as the most informative latent-ranked gene for classifying HNSCC samples. Multi-omic analysis showed promoter hypomethylation without a corresponding expression change, rare mutation, shallow copy-number alteration and no association with survival, and biological interpretation suggests a possible role through Rap1 signalling, MAPK activity and secretion that now requires experimental validation. The study highlights the value of looking beyond conventional differential expression to understand the complex molecular factors that influence cancer behaviour.

432 Data and code availability

433 The head and neck cancer compendium is available from the DeepProfile resource (Qiu et al., 2025); its
 434 GEO accessions are listed in Table 1. The validation datasets are GSE178537 and GSE290057. TCGA-
 435 HNSC data are available from the GDC (<https://portal.gdc.cancer.gov>) and cBioPortal
 436 (<https://www.cbioportal.org>). Code, result tables and figures are available at
 437 <https://github.com/Prokash21/Deep-Neural-Profiling>.

438 Supplementary material



439 **Supplementary Figure S1.** Genome-wide promoter methylation in TCGA-HNSC. (A) Differential methylation of
 440 394,533 CpGs (tumour vs normal). (B) Starburst plot of mean promoter $\Delta\beta$ against expression log2 fold change for
 441 14,066 genes. (C) Overlap of promoter-hypermethylated and downregulated genes. (D) Overlap of promoter-
 442 hypomethylated and upregulated genes.

443 Supplementary Tables S1–S10 (supplementary/Supplementary_Tables_S1-S10.xlsx):

- 444 • **Table S1.** GEO datasets of the DeepProfile head and neck cancer compendium (accession, total samples in
 445 series, platform, title).
- 446 • **Table S2.** Summary of genome-wide differential methylation in TCGA-HNSC.
- 447 • **Table S3.** Summary of the methylation–expression integration.

- **Table S4.** Fisher's exact tests for overlaps between methylation and expression gene sets.
- **Table S5.** Genes with concordant promoter methylation and expression change and significant negative correlation (163 Hyper-Down, 42 Hypo-Up).
- **Table S6.** Tumour versus matched normal differential expression of RAP1GAP2 and RAP1GAP in two cohorts.
- **Table S7.** Copy-number and mutation frequencies of RAP1GAP2 and RAP1GAP in TCGA-HNSC.
- **Table S8.** RAP1GAP2 somatic mutation in TCGA-HNSC.
- **Table S9.** Downstream marker comparison between RAP1GAP2-high and -low tumours.
- **Table S10.** RAP1GAP2 promoter methylation, expression and correlation summary.

Author contributions

J.P.D. conceived the study, performed the analyses and wrote the manuscript. T.H. and P.R. supervised the study and reviewed and edited the manuscript. All authors approved the final version.

Acknowledgements

We thank the DeepProfile authors for making their framework and data publicly available, and the TCGA Research Network and GEO submitters for the data used here.

Competing interests

The authors declare no competing interests.

References

- Abbas M, El-Manzalawy Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC Medical Genomics*. 2020;13:122. <https://doi.org/10.1186/s12920-020-00771-4>
- Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363–1369. <https://doi.org/10.1093/bioinformatics/btu049>
- Behdenna A, Colange M, Haziza J, Gema A, Appé G, Azencott CA, Nordor A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics*. 2023;24:459. <https://doi.org/10.1186/s12859-023-05578-5>
- Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288–295. <https://doi.org/10.1016/j.ygeno.2011.07.007>
- Bos JL, de Rooij J, Reedquist KA. Rap1 signalling: adhering to new models. *Nature Reviews Molecular Cell Biology*. 2001;2(5):369–377. <https://doi.org/10.1038/35073073>
- Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576–582. <https://doi.org/10.1038/nature14129>
- Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*. 2012;2(5):401–404. <https://doi.org/10.1158/2159-8290.CD-12-0095>
- Colaprico A, Silva TC, Olsen C, et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*. 2016;44(8):e71. <https://doi.org/10.1093/nar/gkv1507>
- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Janizek JD, Spiro A, Celik S, et al. PAUSE: principled feature attribution for unsupervised gene expression analysis. *Genome Biology*. 2023;24:81. <https://doi.org/10.1186/s13059-023-02901-4>

Johnson DE, Burtneß B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*. 2020;6:92. <https://doi.org/10.1038/s41572-020-00224-3>

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–127. <https://doi.org/10.1093/biostatistics/kxj037>

Kingma DP, Welling M. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*. 2014. <https://arxiv.org/abs/1312.6114>

Lechner M, Liu J, Masterson L, Fenton TR. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nature Reviews Clinical Oncology*. 2022;19(5):306–327. <https://doi.org/10.1038/s41571-022-00603-7>

Li Z, Fang R, Fang J, He S, Liu T. Functional implications of Rab27 GTPases in cancer. *Cell Communication and Signaling*. 2018;16:44. <https://doi.org/10.1186/s12964-018-0255-9>

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15:550. <https://doi.org/10.1186/s13059-014-0550-8>

Lu CW, Lin SC, Chien CW, et al. Overexpression of pyruvate dehydrogenase kinase 3 increases drug resistance and early recurrence in colon cancer. *American Journal of Pathology*. 2011;179(3):1405–1414. <https://doi.org/10.1016/j.ajpath.2011.05.050>

Mitra RS, Goto M, Lee JS, et al. Rap1GAP promotes invasion via induction of matrix metalloproteinase 9 secretion, which is associated with poor survival in low N-stage squamous cell carcinoma. *Cancer Research*. 2008;68(10):3959–3969. <https://doi.org/10.1158/0008-5472.CAN-07-2755>

Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics*. 2023;39(9):btad547. <https://doi.org/10.1093/bioinformatics/btad547>

Neumüller O, Hoffmeister M, Babica J, Prella C, Gegenbauer K, Smolenski AP. Synaptotagmin-like protein 1 interacts with the GTPase-activating protein Rap1GAP2 and regulates dense granule secretion in platelets. *Blood*. 2009;114(7):1396–1404. <https://doi.org/10.1182/blood-2008-05-155234>

Peterson H, Kolberg L, Raudvere U, Kuzmin I, Vilo J. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research*. 2020;9:ELIXIR-709. <https://doi.org/10.12688/f1000research.24956.2>

Price LS, Hajdo-Milasinovic A, Zhao J, Zwartkruis FJT, Collard JG, Bos JL. Rap1 regulates E-cadherin-mediated cell-cell adhesion. *Journal of Biological Chemistry*. 2004;279(34):35127–35132. <https://doi.org/10.1074/jbc.M404917200>

Qiu W, Dincer AB, Janizek JD, Celik S, Pittet MJ, Naxerova K, Lee SI. Deep profiling of gene expression across 18 human cancers. *Nature Biomedical Engineering*. 2025;9(3):333–355. <https://doi.org/10.1038/s41551-024-01290-8>

Rampášek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*. 2019;35(19):3743–3751. <https://doi.org/10.1093/bioinformatics/btz158>

Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47. <https://doi.org/10.1093/nar/gkv007>

Sabbatini G, Manganaro L. On potential limitations of differential expression analysis with non-linear machine learning models. *EMBnet.journal*. 2023;28:e1035. <https://doi.org/10.14806/ej.28.0.1035>

Schultess J, Danielewski O, Smolenski AP. Rap1GAP2 is a new GTPase-activating protein of Rap1 expressed in human platelets. *Blood*. 2005;105(8):3185–3192. <https://doi.org/10.1182/blood-2004-09-3605>

Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences USA*. 2005;102(43):15545–15550. <https://doi.org/10.1073/pnas.0506580102>

Sun N, Zhao X. Therapeutic implications of FABP4 in cancer: an emerging target to tackle cancer. *Frontiers in Pharmacology*. 2022;13:948610. <https://doi.org/10.3389/fphar.2022.948610>

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017.

- 534 Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent
535 space dimensionalities learns complementary biological representations. *Genome Biology*. 2020;21:109.
536 <https://doi.org/10.1186/s13059-020-02021-3>
- 537 Zhang YL, Wang RC, Cheng K, Ring BZ, Su L. Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer*
538 *Biology and Medicine*. 2017;14(1):90–99. <https://doi.org/10.20892/j.issn.2095-3941.2016.0086>
- 539 Zhang Z, Mitra RS, Henson BS, et al. Rap1GAP inhibits tumor growth in oropharyngeal squamous cell carcinoma.
540 *American Journal of Pathology*. 2006;168(2):585–596. <https://doi.org/10.2353/ajpath.2006.050132>
- 541 Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving
542 large differences. *Bioinformatics*. 2019;35(12):2084–2092. <https://doi.org/10.1093/bioinformatics/bty895>